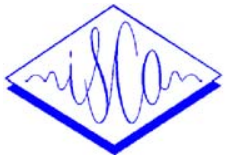

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

COMPARING DIFFERENT MODEL CONFIGURATIONS FOR LANGUAGE IDENTIFICATION USING A PHONOTACTIC APPROACH

D. Matrouf[1,2], *M. Adda-Decker*[1], *J.L. Gauvain*[1], *L. Lamel*[1]
[1]LIMSI-CNRS, BP 133, 91403 Orsay cedex, France
[2] LIA, University of Avignon
{matrouf,madda,gauvain,lamel}@limsi.fr

10.21437/Eurospeech.1999-100

ABSTRACT

In this paper different model configurations for language identification using a phonotactic approach are explored. Identification experiments were carried out on the 11-language telephone speech corpus OGI-TS, containing calls in French, English, German, Spanish, Japanese, Korean, Mandarin, Tamil, Farsi, Hindi, and Vietnamese. Phone sequences output by one or multiple phone recognizers are rescored with language-dependent phonotactic models approximated by phone bigrams. The parameters of different sets of acoustic phone models were estimated using the 4-language IDEAL corpus. Sets of language-specific phonotactic models were trained using the training portion of the OGI-TS CORPUS. Error rates are significantly reduced by combining language-dependent and language-independent acoustic decoders, especially for short segments. A 9.9% LID error rate was obtained on the 11-language task using phonotactic models trained on spontaneous speech data. These results show that the phonotactic approach is relative insensitive to an acoustic mismatch between training and test conditions.

INTRODUCTION

Various information sources can be exploited in order to identify a given language: acoustic, phonetic, phonotactic, lexical, etc. Automatic language identification (LID) may be based on different types and combinations of these information sources. Their modeling requires specific resources, knowledge and corpora, for each language. Acoustic-phonetic and lexical approaches typically make use of language-dependent acoustic phone models, language-dependent phone bigrams, and for a lexical approach, a more or less comprehensive vocabulary for each language [1]. In addition to speech corpora, orthographic and/or phonemic transcripts are needed for model estimation. Phonemic transcripts are commonly obtained by aligning the speech signal with the acoustic phone model graph corresponding to predictable pronunciations of the words in the orthographic transcription. These resources may be difficult or impossible to gather, preventing easy extension to a new language. If such an extension is a required feature for an LID system, a phonotactic approach may be more appropriate. Previous work (see for example, [2, 3]) has demonstrated the interest of a phonotactic approach for LID, especially in dealing with new and/or unknown languages. The acoustic stream of a given language is converted by one (or multiple) phone-based acoustic decoders into one (or multiple) phone label streams which serve during model training as input for phonotactic language model (LM) estimation. During the identification phase, the decoded phone stream serves as input to the phonotactic decoder. This decoder uses the language-dependent phonotactic models to score the input, and a decision module uses the set of language-dependent LM scores to hypothesize the identified language.

In previous work [3] the use of 4 parallel language-dependent (LD) acoustic decoders was compared to the use of one single language-independent (LI) or more accurately one multi-language acoustic decoder. We showed that comparable results could be obtained applying a phonotactic approach to multiple LD phone label streams and to one LI label stream. These results were based on the 4-language (French, English, German, Spanish) IDEAL telephone database. The experiments presented in this paper use the OGI-TS corpus to extend the phonotactic approach to an 11-language task. The use of the OGI data also allows easier comparison with other research.

In this contribution we focus on several aspects of a phonotactic LID system. A major part is devoted to the role of the acoustic decoder providing the phone label streams which are the inputs to the phonotactic decoder which uses the language-specific phonotactic models. Special attention is focused on the use of language-independent acoustic decoders and the parallel use of multiple acoustic decoders. Another aspect concerns the estimation of phone bigrams providing the phonotactic constraints. The influence of training material selection on the phonotactic model accuracy is investigated.

METHOD

The phonotactic approach is particularly useful when labeled speech corpora are not available or are difficult to obtain. Language-specific phonotactic models can be trained using automatically labeled data. The only resources required are one or more acoustic-phonetic decoders. The acoustic signal x of language l is automatically transcribed using a given acoustic-decoder (e.g. of language k) resulting in the phone sequence Φ_k. The resulting phone sequences are then used to train decoder-dependent language-specific phone bigrams as shown in Figure 1.

The acoustic decoder of language k can optionally include a phone bigram language model for the language, where the phone sequence Φ_k is obtained by maximizing $f(x|\Phi,k)\,\Pr(\Phi|k)$. However since we are interested in ap-

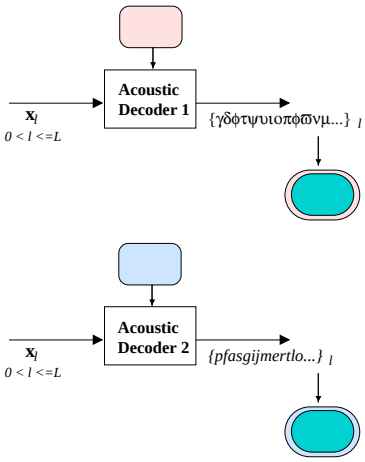

Figure 1: Training of decoder-dependent, language-specific phonotactic models for language l using $K = 2$ acoustic decoders in parallel.

plying the decoder to languages for which acoustic models are not available (ie., in general language $l \neq k$) it may be more appropriate to use only acoustic information, defining Φ_k as follows:

$$\Phi_k = \underset{\Phi}{\operatorname{argmax}}\, f(x|\Phi, k)$$

The phone sequence Φ_k for language l can be used to estimate a language-specific phonotactic model $\Pr(\Phi_k|l)$.

The LID problem can then be viewed as follows:

$$l^* = \underset{l}{\operatorname{argmax}} \Pr(\Phi_k|l)$$

If acoustic decoders are available for K languages, K phone streams $(\Phi_1, \ldots, \Phi_k, \ldots, \Phi_K)$ can be produced in parallel for a given signal x, and corresponding phonotactic models may be estimated. Under the assumption that the K phone streams are independent, the LID decision can be written as follows:

$$l^* = \underset{l}{\operatorname{argmax}} \prod_k \Pr(\Phi_k|l)$$

The corresponding LID system can then be represented as shown in Figure 2 with $K = 2$ decoders and $L = 3$ languages.

EXPERIMENTAL SETUP

In these experiments the test data come from the OGI-TS corpus, and the acoustic models were trained on the 4-language IDEAL corpus. The IDEAL corpus was developed to carry out research in automatic language identification, and contains telephone speech in four-languages (French, British English, German and Castillan Spanish). The corpus is similar in style to the OGI-TS multi-language corpus [6], containing read, elicited and spontaneous speech for each caller, with a larger proportion of read and elicited speech. The spontaneous data accounts for about 15% of the IDEAL corpus. The corpus contains data from about 300 native speakers of each language calling from their home country, of which 250 calls per language were used

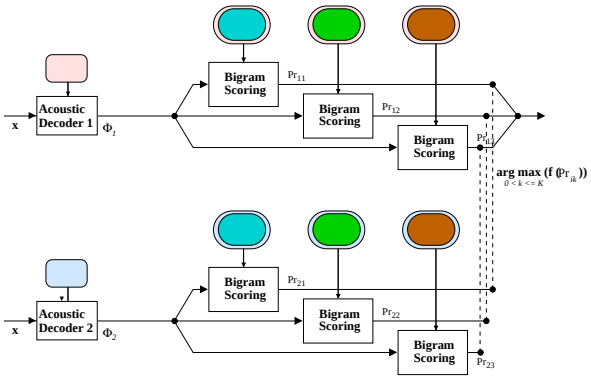

Figure 2: LID system architecture using a phonotactic approach with $K = 2$ decoders and $L = 3$ languages to be identified.

			*#phones (*1000) per decoder*				
Lang	*#calls*	*#hours*	*Fr*	*Eng*	*Ger*	*Sp*	*LI*
Sp	84	2h19'	69	74	85	81	66
Eng	84	2h18'	70	72	86	86	66
Ger	79	2h11'	69	72	83	83	65
Fr	80	2h11'	64	68	80	79	60
Tam	82	2h07'	61	65	75	74	58
Man	80	1h58'	51	54	63	61	48
Viet	75	1h50'	46	49	57	56	42
Far	73	1h52'	53	56	66	64	50
Jap	66	1h44'	52	54	63	61	48
Kor	63	1h34'	43	46	53	52	40
Hin	99	1h18'	35	37	43	42	33

Table 1: For each language of the OGI-TS training set the volume of data is given in terms of acoustic signal duration and the number (*1000) of automatically decoded phone symbols by the $K = 5$ different acoustic decoders (*Fr*, *Eng*, *Ger*, *Sp*, *LI*).

for acoustic model training. This corresponds to over 10 hours of speech per language. Using these data context-independent, gender-independent 3-state acoustic phone models were trained for each language. There are 44, 24, 34 and 47 phone models for British English, Spanish, French and German respectively (not including silence). A language-independent (LI) phone set of 90 units was automatically defined by applying an agglomerative hierarchical clustering algorithm [3] to the 4 language-dependent acoustic phone model sets. The 90 acoustic phone models of the LI set have been trained using all the IDEAL training data. More details about the IDEAL corpus and the acoustic phone models may be found in [4, 1].

The 11-language OGI-TS corpus [6] is used for phonotactic model training and for testing. The 11 languages are French, American English, German, Spanish, Japanese, Korean, Mandarin, Tamil, Farsi, Hindi, Vietnamese. The training and testing configurations roughly correspond to the 1994 NIST evaluation. The test data correspond to the 1994 test set with about 20 calls per language. About 80 calls per language are available for training the phonotactic models. The volume of training data is shown in Table 1 in terms of number of calls and duration. These calls include both spontaneous and elicited speech, with the exception of Hindi for which only spontaneous (story) data were

available. This explains the lower volume of data for this language. In order to use comparable amounts of data for each language, not all of the 164 calls in English were used. Phonotactic models are thus trained on about 2 hours of speech for almost all languages. The number of phone symbols as decoded by the different acoustic-phonetic IDEAL decoders (*Fr*, *Eng*, *Ger*, *Sp*, *LI*) are provided for information.

A 39-dimensional feature vector including the energy and 12 Mel-weighted cepstral parameters, and their first and second order derivatives was used in all experiments. Cepstral mean subtraction is carried out in order to normalize to some extent for varying acoustic channel conditions.

EXPERIMENTAL RESULTS

Two types of results are presented. In the first set of experiments, different acoustic decoding configurations are compared, using the same phonotactic training material in terms of training speech (the phone streams used for LM training obviously vary with the acoustic decoder configuration). Phonotactic models are estimated here using all the training material as described in Table 1. The second part of the experiments focuses on the impact of different training material selections and phonotactic model combinations on LID results.

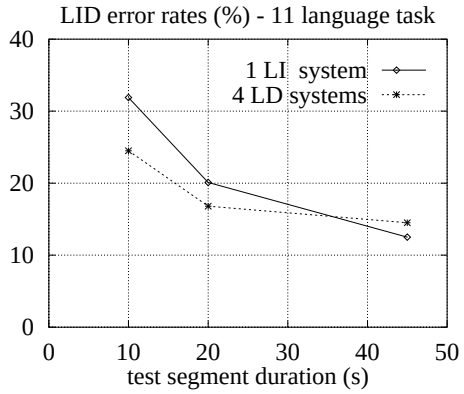

Figure 3: LID error rates on the 11-language OGI NIST eval'94 test set as a function of test segment duration (10s to 45s). Results are given for 4 parallel language-dependent acoustic decoders and 1 language-independent decoder.

Acoustic Decoder Configurations

In previous work exploring phonotactic approaches for LID [3], the use of 4 parallel language-specific acoustic-phonetic decoders was compared with a single language-independent (LI) decoder under both acoustic matched and mismatched (crossed) conditions. Both systems achieved comparable results on 10s chunks of test data, with the LI decoder being slightly better under matched acoustic conditions and slightly worse under mismatched conditions. The experiments described below correspond to the mismatched acoustic condition since the training and test corpora were recorded in independent locations and under different conditions.

Language-dependent vs. language-independent acoustic models: Figure 3 provides results comparing the use of 4 language-dependent decoders in parallel to one language-independent decoder. Whereas for shorter segments (10s, 20s) the 4 parallel decoder configuration provides better results, the situation is reversed on longer segments (45s). Using the single LI decoder, the LID system achieves a 12.5% error rate on the 11-language task.

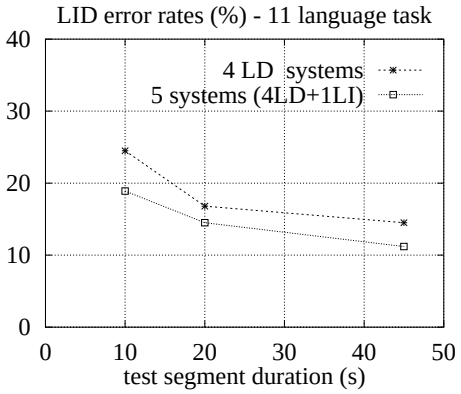

Figure 4: LID error rates on the 11-language OGI NIST eval'94 test set as a function of test segment duration (10s to 45s). Results correspond to 4 LD and 5 (4LD+1LI) decoders in parallel.

The combined use of language-dependent and language-independent acoustic-phonetic decoders is illustrated in Figure 4. The addition of the LI system is seen to provide a relative gain of 10-20% in language identification across test durations. The absolute error rate is reduced from 12.5% to 11.2% on 45s chunks.

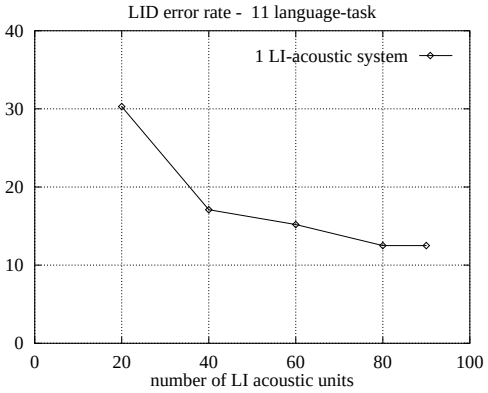

Figure 5: LID error rate on the 45s speech chunks of the 11-language OGI NIST eval'94 test set, using 1 language-independent acoustic decoder. Error rates are shown for different numbers of LI acoustic units.

Language-independent acoustic models: In previous work using the language-independent acoustic phone models, a set of 90 phones was used. The number of classes was manually fixed based on a linguistic analysis of automatically determined phone classes [8]. Figure 5 investigates the impact of the number of language-independent phone classes on LID results. (For language-dependent acoustic models there are typically between 25 amd 50 phone units.) Results on 45s chunks show that the exact number of LI phone models is not critical. A larger number of acoustic models (80 to 90) yields significantly better results than smaller inventories (20 to 40).

	LID err(%)			
bigram LM	1LI phone stream		5 phone streams	
training data	T_{red}	T_{all}	T_{red}	T_{all}
all	14.5	12.5	12.5	11.2
spontaneous	16.5	17.1	14.5	9.9
interpolated	13.2	12.5	11.2	10.5

Table 2: LID error rates with different phonotactic LMs using 45s chunks. T_{red} contains about 60 calls per language, T_{all} on average 80 calls per language. Results are shown for 1 LI acoustic decoder (left) and 5 decoders in parallel (right).

Phonotactic Model Estimation

The phonotactic models are approximated by phone bigrams, which are estimated from the automatically decoded phone streams. To measure the dependence of LID results on the amount of training data, the training material T_{all} (roughly 80 calls per language) was reduced by 25% to about 60 calls per language T_{red}. Three types of language models (LMs) were estimated depending on the type of selected material: *all*, *spontaneous* and *interpolated*. The interpolated LM is obtained from LM_{all} and LM_{spon} using an interpolation coefficient of 0.5. This is equivalent to giving a higher weight to the spontaneous speech as compared to elicited speech.

LID results are shown in Table 2 using 1-LI and 5 (4LD+1LI) acoustic decoders respectively. Reducing the training material from 80 to 60 calls entails an error rate increase in almost all situations. Larger variations in the error rate can be observed with the 1-LI configuration as compared to the 5-decoder configuration. The best results were obtained with the interpolated LMs. Using T_{all} and the 5 acoustic decoder configuration the spontaneous LM achieves the best result LID error rate of 9.9% .

DISCUSSION

Experiments have been carried out with a phonotactic-based approach LID system on an 11-language task (OGI-TS corpus). The main advantage of such an approach is its straightforward extension to new languages. The phonotactic approach generally requires longer test segments to obtain optimal results as compared to acoustic-phonetic approaches, but it is less channel-sensitive [3]. Using a phonotactic approach LID results significantly improve when the test segment length goes from 10s to 45s. Combining several acoustic decoders in parallel yields important gains especially on shorter test segments.

Direct comparison between one single language-independent decoder and 4 language-dependent decoders showed that the LI configuration performs better on longer test segments (45s), while the parallel architecture is more effective on shorter segments (10s). Adding the language-independent decoder to the 4 language-dependent decoders resulted in a consistent gain in all configurations. We have also carried out similar experiments by varying the number of language-dependent decoders. More decoders did not systematically improve LID performance: a configuration using 2 LD decoders in parallel can perform as well as one with 4 LD decoders in parallel.

A study concerning the number of LI models showed that the exact number of acoustic phone models is not critical. Best performances were observed with 80 to 90 acoustic phone models. An interesting outcome of our experiments concerns the robustness of the phonotactic approach with respect to acoustically mismatched conditions for training and test. Our results show that state-of-the-art LID results [7, 2, 5] may be obtained with acoustic models trained on independent corpora. In our experimental setup the acoustic models were trained on a 4-language corpus containing calls from 4 different European countries[4].

The results using the different phonotactic model sets show that the volume of training data in our experimental setup remains critical for reliably estimating the parameters of the phonotactic model. Large variations were also observed by varying the type of training data. These variations were reduced when moving from a single LI acoustic decoder configuration to a multiple decoder configuration. The multiple decoder configuration seems to be of particular interest in at least two situations: when the test segments are relatively short and when there is limited phonotactic model training data. We are presently investigating the impact of adapting acoustic models to the test condition and improving the LID decision module. In preliminary experiments using the LDA (linear discriminant analysis) technique an LID error rate of 8.5% was achieved on 45s chunks with a 5 parallel acoustic decoder configuration and OGI-adapted acoustic models.

REFERENCES

[1] D. Matrouf, M. Adda-Decker, L. Lamel, J.L. Gauvain, "Language Identification Incorporating Lexical Information," *ICSLP'98*, pp. 181-184, Sydney, Dec. 1998.

[2] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. on SAP*, **4**(1), pp. 31-34, Jan. 1996.

[3] C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, L. Lamel, "Language Identification with Language-Independent Acoustic Models", *Eurospeech'97*, pp. 5-8, Rhodes, Sep. 1997.

[4] L. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy, J.J. Gangolf, J.L. Gauvain, "A Multilingual Corpus for Language Identification," *1st International Conference on Language Resources and Evaluation*, **1**, pp. 1115-1122, Granada, May 1998.

[5] M.A. Zissman, "Predicting, Diagnosing and Improving Automatic Language Identification Performance," *Eurospeech'97*, Rhodes, Sep. 1997.

[6] Y.K. Muthusamy, R.A. Cole, B.T. Oshika, "The OGI Multi-Language Telephone Speech Corpus," *ICSLP'92*, pp. 895-898, Banff, Oct. 1992.

[7] J. Navratil, W. Zühlke, "An Efficient Phonotactic-Acoustic System for Language Identification," *ICASSP'98*, pp. 781-784, Seattle, May 1998.

[8] P. Boula de Mareüil, C. Corredor Ardoy, M. Adda-Decker, "Multi-lingual automatic phoneme clustering", *14th Int. Conf. on Phonetic Science, ICPhS-99*, August 1999.